

Orthogonal Matching Pursuit: A Brownian Motion Analysis

Alyson K. Fletcher and Sundeep Rangan

Abstract—A well-known analysis of Tropp and Gilbert shows that orthogonal matching pursuit (OMP) can recover a k -sparse n -dimensional real vector from $m = 4k \log(n)$ noise-free linear measurements obtained through a random Gaussian measurement matrix with a probability that approaches one as $n \rightarrow \infty$. This work strengthens this result by showing that a lower number of measurements, $m = 2k \log(n - k)$, is in fact sufficient for asymptotic recovery. More generally, when the sparsity level satisfies $k_{\min} \leq k \leq k_{\max}$ but is unknown, $m = 2k_{\max} \log(n - k_{\min})$ measurements is sufficient. Furthermore, this number of measurements is also sufficient for detection of the sparsity pattern (support) of the vector with measurement errors provided the signal-to-noise ratio (SNR) scales to infinity. The scaling $m = 2k \log(n - k)$ exactly matches the number of measurements required by the more complex lasso method for signal recovery with a similar SNR scaling.

Index Terms—compressed sensing, detection, lasso, orthogonal matching pursuit, random matrices, sparse approximation, sparsity, subset selection

I. INTRODUCTION

Suppose $\mathbf{x} \in \mathbb{R}^n$ is a sparse vector, meaning its number of nonzero entries k is smaller than n . The *support* of \mathbf{x} is the locations of the nonzero entries and is sometimes called its *sparsity pattern*. A common sparse estimation problem is to infer the sparsity pattern of \mathbf{x} from linear measurements of the form

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{w}, \quad (1)$$

where $\mathbf{A} \in \mathbb{R}^{m \times n}$ is a known measurement matrix, $\mathbf{y} \in \mathbb{R}^m$ represents a vector of measurements and $\mathbf{w} \in \mathbb{R}^m$ is a vector of measurement errors (noise).

Sparsity pattern detection and related sparse estimation problems are classical problems in nonlinear signal processing and arise in a variety of applications including wavelet-based image processing [1] and statistical model selection in linear regression [2]. There has also been considerable recent interest in sparsity pattern detection in the context of *compressed sensing*, which focuses on large random measurement matrices \mathbf{A} [3]–[5]. It is this scenario with random measurements that will be analyzed here.

Optimal subset recovery is NP-hard [6] and usually involves searches over all the $\binom{n}{k}$ possible support sets of \mathbf{x} . Thus, most

attention has focused on approximate methods. One simple and popular approximate algorithm is orthogonal matching pursuit (OMP) [7]–[9]. OMP is a greedy method that identifies the location of one nonzero entry of \mathbf{x} at a time. A version of the algorithm will be described in detail below in Section II. The best known analysis of the detection performance of OMP for large random matrices is due to Tropp and Gilbert [10], [11]. Among other results, Tropp and Gilbert show that when \mathbf{A} has i.i.d. Gaussian entries, the measurements are noise-free ($\mathbf{w} = 0$), and the number of measurements scales as

$$m \geq (1 + \delta)4k \log(n) \quad (2)$$

for some $\delta > 0$, the OMP method will recover the correct sparse pattern of \mathbf{x} with a probability that approaches one as n and $k \rightarrow \infty$. The analysis uses a deterministic sufficient condition for success on the matrix \mathbf{A} based on a greedy selection ratio introduced in [12]. A similar deterministic condition on \mathbf{A} was presented in [13], and a condition using the restricted isometry property was given in [14].

Numerical experiments reported in [10] suggest that a smaller number of measurements than (2) may be sufficient for asymptotic recovery with OMP. Specifically, the experiments suggest that the constant 4 can be reduced to 2.

Our main result, Theorem 1 below, does a bit better than proving this conjecture. We show that the scaling in measurements

$$m \geq (1 + \delta)2k \log(n - k) \quad (3)$$

is sufficient for asymptotic reliable recovery with OMP provided both $n - k$ and $k \rightarrow \infty$. Theorem 1 goes further by allowing uncertainty in the sparsity level k .

We also improve upon the Tropp–Gilbert analysis by accounting for the effect of the noise \mathbf{w} . While the Tropp–Gilbert analysis requires that the measurements are noise-free, we show that the scaling (3) is also sufficient when there is noise \mathbf{w} , provided the signal-to-noise ratio (SNR) goes to infinity.

The main significance of the new scaling (3) is that it exactly matches the conditions for sparsity pattern recovery using the well-known lasso method. The lasso method, which will be described in detail in Section IV, is based on a convex relaxation of the optimal detection problem. The best analysis of sparsity pattern recovery with lasso is due to Wainwright [15], [16]. He showed in [15] that under a similar high SNR assumption, the scaling (3) in number of measurements is both necessary and sufficient for asymptotic

This work was supported in part by a University of California President's Postdoctoral Fellowship. The material in this paper was presented in part at the Conference on Neural Information Processing Systems, Vancouver, BC, Canada, December 2009.

A. K. Fletcher (email: alyson@eecs.berkeley.edu) is with the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley.

S. Rangan (email: srangan@poly.edu) is with the Polytechnic Institute of New York University, Brooklyn, NY.

reliable sparsity pattern detection.¹ The lasso method is often more complex than OMP, but it is widely believed to offset this disadvantage with superior performance [10]. Our results show that, at least for sparsity pattern recovery under our asymptotic assumptions, OMP performs at least as well as lasso.² Hence, the additional complexity of lasso for these problems may not be warranted.

Neither lasso nor OMP is the best known approximate algorithm for sparsity pattern recovery. For example, where there is no noise in the measurements, the lasso minimization (15) can be replaced by

$$\hat{\mathbf{x}} = \arg \min_{\mathbf{v} \in \mathbb{R}^n} \|\mathbf{v}\|_1, \text{ s.t. } \mathbf{y} = \mathbf{A}\mathbf{v}.$$

A well-known analysis due to Donoho and Tanner [17] shows that, for i.i.d. Gaussian measurement matrices, this minimization will recover the correct vector with

$$m \asymp 2k \log(n/m) \quad (4)$$

when $k \ll n$. This scaling is fundamentally better than the scaling (3) achieved by OMP and lasso.

There are also several variants of OMP that have shown improved performance. The CoSaMP algorithm of Needell and Tropp [18] and subspace pursuit algorithm of Dai and Milenkovic [19] achieve a scaling similar to (4). Other variants of OMP include the stagewise OMP [20] and regularized OMP [21], [22]. Indeed with the recent interest in compressed sensing, there is now a wide range of promising algorithms available. We do not claim that OMP achieves the best performance in any sense. Rather, we simply intend to show that both OMP and lasso have similar performance in certain scenarios.

Our proof of (3) follows along the same lines as Tropp and Gilbert's proof of (2), but with two key differences. First, we account for the effect of the noise by separately considering its effect in the "true" subspace and its orthogonal complement. Second and more importantly, we address the "nasty independence issues" noted by Tropp and Gilbert [10] by providing a tighter bound on the maximum correlation of the incorrect vectors. Specifically, in each iteration of the OMP algorithm, there are $n - k$ possible incorrect vectors that the algorithm can choose. Since the algorithm runs for k iterations, there are total of $k(n - k)$ possible error events. The Tropp and Gilbert proof bounds the probability of these error events with a union bound, essentially treating them as statistically independent. However, here we show that energies on any one of the incorrect vectors across the k iterations are correlated. In fact, they are precisely described by samples of a certain normalized Brownian motion. Exploiting this correlation we show that the tail bound on error probability grows as $n - k$, not $k(n - k)$, independent events.

¹Sufficient conditions under weaker conditions on the SNR are more subtle [16]: the scaling of SNR with n determines the sequences of regularization parameters for which asymptotic almost sure success is achieved, and the regularization parameter sequence affects the sufficient number of measurements.

²Recall that our result is a sufficient condition for success whereas the matching condition for lasso is both necessary and sufficient.

The outline of the remainder of this paper is as follows. Section II describes the OMP algorithm. Our main result, Theorem 1, is stated in Section III. A comparison to lasso is provided in Section IV, and we suggest some future problems in Section VII. The proof of the main result is somewhat long and given in the Section VIII. The main result was first reported in [23].

II. ORTHOGONAL MATCHING PURSUIT

To describe the algorithm, suppose we wish to determine the vector \mathbf{x} from a vector \mathbf{y} of the form (1). Let

$$I_{\text{true}} = \{j : x_j \neq 0\}, \quad (5)$$

which is the support of the vector \mathbf{x} . The set I_{true} will also be called the *sparsity pattern*. Let $k = |I_{\text{true}}|$, which is the number of nonzero entries of \mathbf{x} . The OMP algorithm produces a sequence of estimates $\hat{I}(t)$, $t = 0, 1, 2, \dots$, of the sparsity pattern I_{true} , adding one index at a time. In the description below, let \mathbf{a}_j denote the j th column of \mathbf{A} .

Algorithm 1 (Orthogonal Matching Pursuit): Given a vector $\mathbf{y} \in \mathbb{R}^m$, a measurement matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, and threshold level $\mu > 0$, compute an estimate \hat{I}_{OMP} of the sparsity pattern of \mathbf{x} as follows:

- 1) Initialize $t = 0$ and $\hat{I}(t) = \emptyset$.
- 2) Compute $\mathbf{P}(t)$, the projection operator onto the orthogonal complement of the span of $\{\mathbf{a}_i, i \in \hat{I}(t)\}$.
- 3) For each j , compute

$$\rho(t, j) = \frac{|\mathbf{a}_j' \mathbf{P}(t) \mathbf{y}|^2}{\|\mathbf{P}(t) \mathbf{y}\|^2},$$

and let

$$[\rho^*(t), i^*(t)] = \max_{j=1, \dots, n} \rho(t, j), \quad (6)$$

where $\rho^*(t)$ is the value of the maximum and $i^*(t)$ is an index that achieves the maximum.

- 4) If $\rho^*(t) > \mu$, set $\hat{I}(t+1) = \hat{I}(t) \cup \{i^*(t)\}$. Also, increment $t = t + 1$ and return to step 2.
- 5) Otherwise stop. The final estimate of the sparsity pattern is $\hat{I}_{\text{OMP}} = \hat{I}(t)$.

Note that since $\mathbf{P}(t)$ is the projection onto the orthogonal complement of the span of $\{\mathbf{a}_j, j \in \hat{I}(t)\}$, for all $j \in \hat{I}(t)$ we have $\mathbf{P}(t) \mathbf{a}_j = 0$. Hence, $\rho(t, j) = 0$ for all $j \in \hat{I}(t)$, and therefore the algorithm will not select the same vector twice.

The algorithm above only provides an estimate, \hat{I}_{OMP} , of the sparsity pattern of I_{true} . Using \hat{I}_{OMP} , one can estimate the vector \mathbf{x} in a number of ways. For example, one can take the least-squares estimate,

$$\hat{\mathbf{x}} = \arg \min \|\mathbf{y} - \mathbf{A}\mathbf{v}\|^2 \quad (7)$$

where the minimization is over all vectors \mathbf{v} such $v_j = 0$ for all $j \notin \hat{I}_{\text{OMP}}$. The estimate $\hat{\mathbf{x}}$ is the projection of the noisy vector \mathbf{y} onto the space spanned by the vectors \mathbf{a}_i with i in the sparsity pattern estimate \hat{I}_{OMP} . This paper only analyzes the sparsity pattern estimate \hat{I}_{OMP} itself, and not the vector estimate $\hat{\mathbf{x}}$.

III. ASYMPTOTIC ANALYSIS

We analyze the OMP algorithm in the previous section under the following assumptions.

Assumption 1: Consider a sequence of sparse recovery problems, indexed by the vector dimension n . For each n , let $\mathbf{x} \in \mathbb{R}^n$ be a deterministic vector. Also assume:

- (a) The sparsity level $k = k(n)$ (i.e., number of nonzero entries in \mathbf{x}) satisfies

$$k(n) \in [k_{\min}(n), k_{\max}(n)] \quad (8)$$

for some deterministic sequences $k_{\min}(n)$ and $k_{\max}(n)$ with $k_{\min}(n) \rightarrow \infty$ as $n \rightarrow \infty$ and $k_{\max}(n) < n/2$ for all n .

- (b) The number of measurements $m = m(n)$ is a deterministic sequence satisfying

$$m \geq (1 + \delta)2k_{\max} \log(n - k_{\min}) \quad (9)$$

for some $\delta > 0$.

- (c) The minimum component power x_{\min}^2 satisfies

$$\lim_{n \rightarrow \infty} kx_{\min}^2 = \infty, \quad (10)$$

where

$$x_{\min} = \min_{j \in I_{\text{true}}} |x_j| \quad (11)$$

is the magnitude of the smallest nonzero entry of \mathbf{x} .

- (d) The powers of the vectors $\|\mathbf{x}\|^2$ satisfy

$$\lim_{n \rightarrow \infty} \frac{1}{(n - k)^\epsilon} \log(1 + \|\mathbf{x}\|^2) = 0 \quad (12)$$

for all $\epsilon > 0$.

- (e) The vector \mathbf{y} is a random vector generated by (1) where \mathbf{A} and \mathbf{w} have i.i.d. Gaussian entries with zero mean and variance $1/m$.

Assumption 1(a) provides a range on the sparsity level k . As we will see below in Section V, bounds on this range are necessary for proper selection of the threshold level $\mu > 0$.

Assumption 1(b) is the scaling law on the number of measurements that we will show is sufficient for asymptotically reliable recovery. In the special case when k is known so that $k_{\max} = k_{\min} = k$, we obtain the simpler scaling law

$$m \geq (1 + \delta)2k \log(n - k). \quad (13)$$

We have contrasted this scaling law with the Tropp–Gilbert scaling law (2) in Section I. We will also compare it to the scaling law for lasso in Section IV.

Assumption 1(c) is critical and places constraints on the smallest component magnitude. The importance of the smallest component magnitude in the detection of the sparsity pattern was first recognized by Wainwright [15], [16], [24]. Also, as discussed in [25], the condition requires that signal-to-noise ratio (SNR) goes to infinity. Specifically, if we define the SNR as

$$\text{SNR} = \frac{\mathbf{E}\|\mathbf{Ax}\|^2}{\mathbf{E}\|\mathbf{w}\|^2},$$

then under Assumption 1(e) it can be easily checked that

$$\text{SNR} = \|\mathbf{x}\|^2. \quad (14)$$

Since \mathbf{x} has k nonzero entries, $\|\mathbf{x}\|^2 \geq kx_{\min}^2$, and therefore condition (10) requires that $\text{SNR} \rightarrow \infty$. For this reason, we will call our analysis of OMP a high-SNR analysis. The analysis of OMP with SNR that remains bounded above is an interesting open problem.

Assumption (d) is technical and simply requires that the SNR does not grow too quickly with n . Note that even if $\text{SNR} = O(k^\alpha)$ for any $\alpha > 0$, Assumption 1(d) will be satisfied.

Assumption 1(e) states that our analysis concerns large Gaussian measurement matrices \mathbf{A} and Gaussian noise \mathbf{w} .

Our main result is as follows.

Theorem 1: Under Assumption 1, there exists a sequence of threshold levels $\mu = \mu(n)$ such that the OMP method in Algorithm 1 will asymptotically detect the correct sparsity pattern in that

$$\lim_{n \rightarrow \infty} \Pr(\hat{I}_{\text{OMP}} \neq I_{\text{true}}) = 0.$$

Moreover, the threshold levels μ can be selected simply as a function of k_{\min} , k_{\max} , n , m and δ .

Theorem 1 provides our main scaling law for OMP. The proof is given in Section VIII.

IV. COMPARISON TO LASSO PERFORMANCE

It is useful to compare the scaling law (13) to the number of measurements required by the widely-used lasso method described for example in [26]. The lasso method finds an estimate for the vector \mathbf{x} in (1) by solving the quadratic program

$$\hat{\mathbf{x}} = \arg \min_{\mathbf{v} \in \mathbb{R}^n} \|\mathbf{y} - \mathbf{Av}\|^2 + \mu \|\mathbf{v}\|_1, \quad (15)$$

where $\mu > 0$ is an algorithm parameter that trades off the prediction error with the sparsity of the solution. Lasso is sometimes referred to as basis pursuit denoising [27]. While the optimization (15) is convex, the running time of lasso is significantly longer than OMP unless \mathbf{A} has some particular structure [10]. However, it is generally believed that lasso has superior performance.

The best analysis of lasso for sparsity pattern recovery for large random matrices is due to Wainwright [15], [16]. There, it is shown that with an i.i.d. Gaussian measurement matrix and white Gaussian noise, the condition (13) is *necessary* for asymptotic reliable detection of the sparsity pattern. In addition, under the condition (10) on the minimum component magnitude, the scaling (13) is also sufficient. We thus conclude that OMP requires an identical scaling in the number of measurements to lasso. Therefore, at least for sparsity pattern recovery from measurements with large random Gaussian measurement matrices and high SNR, there is no additional performance improvement with the more complex lasso method over OMP.

V. THRESHOLD SELECTION AND STOPPING CONDITIONS

In many problems, the sparsity level k is not known *a priori* and must be detected as part of the estimation process. In OMP, the sparsity level of the estimate vector is precisely the number

of iterations conducted before the algorithm terminates. Thus, reliable sparsity level estimation requires a good stopping condition.

When the measurements are noise-free and one is concerned only with exact signal recovery, the optimal stopping condition is simple: the algorithm should simply stop whenever there is no more error; that is, $\rho^*(t) = 0$ in (6). However, with noise, selecting the correct stopping condition requires some care. The OMP method as described in Algorithm 1 uses a stopping condition based on testing if $\rho^*(t) > \mu$ for some threshold μ .

One of the appealing features of Theorem 1 is that it provides a simple sufficient condition under which this threshold mechanism will detect the correct sparsity level. Specifically, Theorem 1 provides a range $k \in [k_{\min}, k_{\max}]$ under which there exists a threshold such that the OMP algorithm will terminate in the correct number of iterations. The larger the number of measurements m , the wider one can make the range $[k_{\min}, k_{\max}]$. The formula for the threshold level is given later in (22).

In practice, one may deliberately want to stop the OMP algorithm with fewer iterations than the “true” sparsity level. As the OMP method proceeds, the detection becomes less reliable and it is sometimes useful to stop the algorithm whenever there is a high chance of error. Stopping early may miss some small entries, but it may result in an overall better estimate by not introducing too many erroneous entries or entries with too much noise. However, since our analysis is only concerned with exact sparsity pattern recovery, we do not consider this type of stopping condition.

VI. NUMERICAL SIMULATIONS

To verify the above analysis, we simulated the OMP algorithm with fixed signal dimension $n = 100$ and different sparsity levels k , numbers of measurements m , and randomly-generated vectors \mathbf{x} .

In the first experiment, $\mathbf{x} \in \mathbb{R}^n$ was generated with k randomly placed nonzero values, with all the nonzero entries having the same magnitude $|x_j| = C$ for some $C > 0$. Following Assumption 1(e), the measurement matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ and noise vector $\mathbf{w} \in \mathbb{R}^m$ were generated with i.i.d. $\mathcal{N}(0, 1/m)$ entries. Using (14) and the fact that \mathbf{x} has k nonzero entries with power C^2 , the SNR is given by

$$\text{SNR} = \|\mathbf{x}\|^2 = kC^2,$$

so the SNR can be controlled by varying C .

Fig. 1 plots the probability that the OMP algorithm incorrectly detected the sparsity pattern for different values of k and m . The probability is estimated with 1000 Monte Carlo simulations per (k, m) pair. For each k and m , the threshold level μ was selected as the one with the lowest probability of error, assuming, of course, that the same μ is used across all 1000 Monte Carlo runs.

The solid curve in Fig. 1 is the theoretical number of measurements in (13) from Theorem 1 that guarantees exact sparsity recovery. The formula is theoretically valid as $n \rightarrow \infty$ and $\text{SNR} \rightarrow \infty$. At finite problem sizes, the probability of error for m satisfying (13) will be nonzero. However, Fig. 1 shows

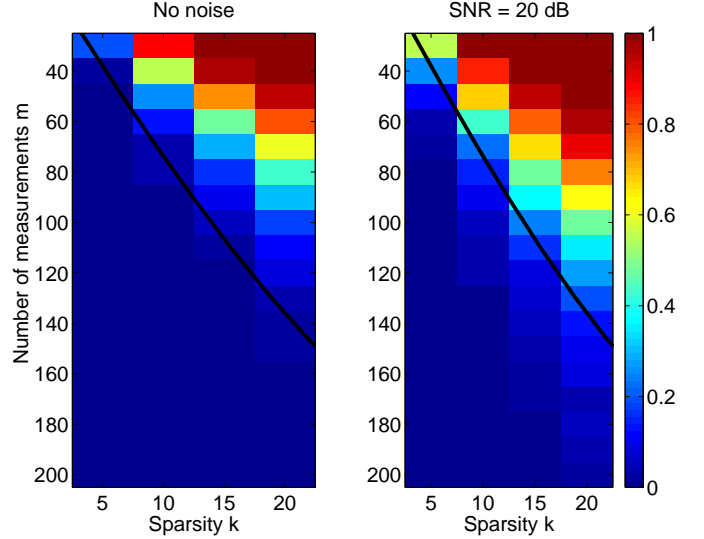


Fig. 1. OMP performance prediction. The colored bars show the probability of sparsity pattern misdetection based on 1000 Monte Carlo simulations of the OMP algorithm. The signal dimension is fixed to $n = 100$ and the error probability is plotted against the number of measurements m and sparsity level k . The solid black curve shows the theoretical number of measurements $m = 2k \log(n - k)$ sufficient for asymptotic reliable detection.

that for the problem size in the simulation, the probability of error for OMP is indeed low for values of m greater than the theoretical level. When there is no noise (i.e. $\text{SNR} = \infty$), the probability of error is between 3 and 5% for most values of k . When the SNR is 20 dB, the probability of error is between 15 and 20%. In either case, the formula provides a reasonable prediction of the threshold in the number of measurements at which the OMP method succeeds.

Theorem 1 is only a *sufficient condition*. It is possible that for some \mathbf{x} , OMP could require a number of measurements less than predicted by (13). That is, the number of measurements (13) may not be *necessary*.

To illustrate such a case, we consider vectors with a nonzero dynamic range of component magnitudes. Fig. 2 shows the probability of sparsity pattern detection as a function of m for vectors \mathbf{x} with different dynamic ranges. Specifically, the k nonzero entries of \mathbf{x} were chosen to have powers uniformly distributed in a range of 0, 10 and 20 dB. In this simulation, we used $k = 20$ and $n = 100$, so the sufficient condition predicted by (13) is $m \approx 136$. When the dynamic range is 0 dB, all the nonzero entries have equal magnitude, and the probability of error at the value $m = 136$ is approximately 3%. However, with a dynamic range of 10 dB, the same probability of error can be achieved with $m \approx 105$ measurements, a value significantly below the sufficient condition in (13). With a dynamic range of 20 dB, the number of measurements decreases further to $m \approx 75$.

This possible benefit of dynamic range in OMP-like algorithms has been observed in [28], [29] and in sparse Bayesian learning [30], [31]. A valuable line of future research would be to see if this benefit can be quantified. That is, it would be useful to develop a sufficient condition tighter than (13) that accounts for the dynamic range of the signals.

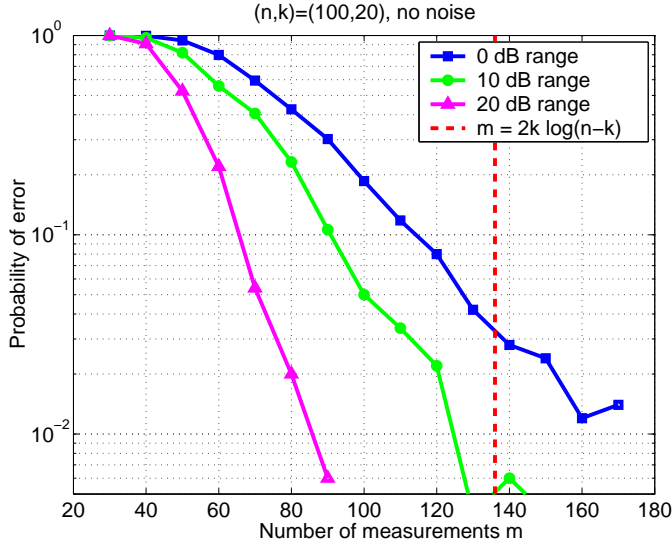


Fig. 2. OMP performance and dynamic range. Plotted is the probability of sparsity pattern detection as a function of the number of measurements for random vectors \mathbf{x} with various dynamic ranges. In all cases, $n = 100$, $k = 20$ and $\text{SNR} = \infty$.

VII. CONCLUSIONS AND FUTURE WORK

We have provided an improved scaling law on the number of measurements for asymptotic reliable sparsity pattern detection with OMP. Most importantly, the scaling law exactly matches the scaling needed by lasso under similar conditions.

However, much about the performance of OMP is still not fully understood. Most importantly, our analysis is limited to high SNR. It would be interesting to see if reasonable sufficient conditions can be derived for finite SNR as well. Also, our analysis has been restricted to exact sparsity pattern recovery. However, in many problems, especially with noise, it is not necessary to detect every element in the sparsity pattern. It would be useful if partial support recovery results such as those in [32]–[34] can be obtained for OMP.

VIII. PROOF OF THEOREM 1

A. Proof Outline

The main difficulty in analyzing OMP is the statistical dependencies between iterations in the OMP algorithm. Following along the lines of the Tropp–Gilbert proof in [10], we avoid these difficulties by considering the following alternate “genie” algorithm. A similar alternate algorithm is analyzed in [28] as well.

- 1) Initialize $t = 0$ and $I_{\text{true}}(t) = \emptyset$.
- 2) Compute $\mathbf{P}_{\text{true}}(t)$, the projection operator onto the orthogonal complement of the span of $\{\mathbf{a}_i, i \in I_{\text{true}}(t)\}$.
- 3) For all $j = 1, \dots, n$, compute

$$\rho_{\text{true}}(t, j) = \frac{|\mathbf{a}'_j \mathbf{P}_{\text{true}}(t) \mathbf{y}|^2}{\|\mathbf{P}_{\text{true}}(t) \mathbf{y}\|^2}, \quad (16)$$

and let

$$[\rho_{\text{true}}^*(t), i^*(t)] = \max_{j \in I_{\text{true}}} \rho_{\text{true}}(t, j). \quad (17)$$

- 4) If $t < k$, set $I_{\text{true}}(t+1) = I_{\text{true}}(t) \cup \{i^*(t)\}$. Increment $t = t + 1$ and return to step 2.
- 5) Otherwise stop. The final estimate of the sparsity pattern is $I_{\text{true}}(k)$.

This “genie” algorithm is identical to the regular OMP method in Algorithm 1, except that it runs for precisely k iterations as opposed to using a threshold μ for the stopping condition. Also, in the maximization in (17), the genie algorithm searches over only the correct indices $j \in I_{\text{true}}$. Hence, this genie algorithm can never select an incorrect index $j \notin I_{\text{true}}$. Also, as in the regular OMP algorithm, the genie algorithm will never select the same vector twice for almost all vectors \mathbf{y} . Therefore, after k iterations, the genie algorithm will have selected all the k indices in I_{true} and terminate with correct sparsity pattern estimate

$$I_{\text{true}}(k) = I_{\text{true}}$$

with probability one.

The reason to consider the sequences $\mathbf{P}_{\text{true}}(t)$ and $I_{\text{true}}(t)$ instead of $\mathbf{P}(t)$ and $\hat{I}(t)$ is that the quantities $\mathbf{P}_{\text{true}}(t)$ and $I_{\text{true}}(t)$ depend only on the vector \mathbf{y} and the columns \mathbf{a}_j for $j \in I_{\text{true}}$. The vector \mathbf{y} also only depends on \mathbf{a}_j for $j \in I_{\text{true}}$ and the noise vector \mathbf{w} . Hence, $\mathbf{P}_{\text{true}}(t)$ and $I_{\text{true}}(t)$ are statistically independent of all the columns \mathbf{a}_j , $j \notin I_{\text{true}}$. This property will be essential in bounding the “false alarm” probability to be defined shortly.

Now, a simple induction argument shows that if

$$\min_{t=0, \dots, k-1} \max_{j \in I_{\text{true}}} \rho_{\text{true}}(t, j) > \mu, \quad (18a)$$

$$\max_{t=0, \dots, k} \max_{j \notin I_{\text{true}}} \rho_{\text{true}}(t, j) < \mu, \quad (18b)$$

then the regular OMP algorithm, Algorithm 1, will terminate in k iterations. Moreover, for all t , the OMP algorithm will output $\mathbf{P}(t) = \mathbf{P}_{\text{true}}(t)$, $\hat{I}(t) = I_{\text{true}}(t)$, and $\rho(t, j) = \rho_{\text{true}}(t, j)$ for all t and j . This will in turn result in the OMP algorithm detecting the correct sparsity pattern

$$\hat{I}_{\text{OMP}} = I_{\text{true}}.$$

So, we need to show that the two events in (18a) and (18b) occur with high probability.

To this end, define the following two probabilities:

$$p_{\text{MD}} = \Pr \left(\max_{t=0, \dots, k-1} \min_{j \in I_{\text{true}}} \rho_{\text{true}}(t, j) \leq \mu \right) \quad (19)$$

$$p_{\text{FA}} = \Pr \left(\max_{t=0, \dots, k} \max_{j \notin I_{\text{true}}} \rho_{\text{true}}(t, j) \geq \mu \right) \quad (20)$$

Both probabilities are implicitly functions of n . The first term, p_{MD} , can be interpreted as a “missed detection” probability, since it corresponds to the event that the maximum correlation energy $\rho_{\text{true}}(t, j)$ on the correct vectors $j \in I_{\text{true}}$ falls below the threshold. We call the second term p_{FA} the “false alarm” probability since it corresponds to the maximum energy on one of the “incorrect” indices $j \notin I_{\text{true}}$ exceeding the threshold.

The above arguments show that

$$\Pr \left(\hat{I}_{\text{OMP}} \neq I_{\text{true}} \right) \leq p_{\text{MD}} + p_{\text{FA}}.$$

So we need to show that there exists a sequence of thresholds $\mu = \mu(n) > 0$, such that $p_{\text{MD}} \rightarrow 0$ and $p_{\text{FA}} \rightarrow 0$ as $n \rightarrow \infty$. We will define the threshold level in Section VIII-B. Sections VIII-C and VIII-D then prove that $p_{\text{MD}} \rightarrow 0$ with this threshold. The difficult part of the proof is to show $p_{\text{FA}} \rightarrow 0$. This part is proven in Section VIII-G after some preliminary results in Sections VIII-E and VIII-F.

B. Threshold Selection

We will first select the threshold sequence $\mu(n)$. Given $\delta > 0$ in (9), let $\epsilon > 0$ such that

$$\frac{1+\delta}{1+\epsilon} \geq 1+\epsilon. \quad (21)$$

Then, define the threshold level

$$\mu = \mu(n) = \frac{2(1+\epsilon)}{m} \log(n - k_{\min}). \quad (22)$$

Observe that since $k \geq k_{\min}$, (22) implies that

$$\mu \geq \frac{2(1+\epsilon)}{m} \log(n - k). \quad (23)$$

Also, since $k \leq k_{\max}$, (9), (21) and (22) show that

$$\mu \leq \frac{1}{(1+\epsilon)k}. \quad (24)$$

C. Decomposition Representation and Related Bounds

To bound the missed detection probability, it is easiest to analyze the OMP algorithm in two separate subspaces: the span of the vectors $\{\mathbf{a}_j, j \in I_{\text{true}}\}$, and its orthogonal complement. This subsection defines some notation for this orthogonal decomposition and proves some simple bounds. The actual limit of the missed detection probability will then be evaluated in the next subsection, Section VIII-D.

Assume without loss of generality $I_{\text{true}} = \{1, 2, \dots, k\}$, so that the vector \mathbf{x} is supported on the first k elements. Let Φ be the $m \times k$ matrix formed by the k correct columns:

$$\Phi = [\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_k].$$

Also, let $\mathbf{x}_{\text{true}} = [x_1, x_2, \dots, x_k]'$ be the vector of the k nonzero entries so that

$$\mathbf{A}\mathbf{x} = \Phi\mathbf{x}_{\text{true}}. \quad (25)$$

Now rewrite the noise vector \mathbf{w} as

$$\mathbf{w} = \Phi\mathbf{v} + \mathbf{w}^\perp \quad (26)$$

where

$$\mathbf{v} = (\Phi'\Phi)^{-1}\Phi'\mathbf{w}, \quad \mathbf{w}^\perp = \mathbf{w} - \Phi\mathbf{v}. \quad (27)$$

The vectors $\Phi\mathbf{v}$ and \mathbf{w}^\perp are, respectively, the projections of the noise vector \mathbf{w} onto the k -dimensional range space of Φ and its orthogonal complement. Combining (25) with (26), we can rewrite (1) as

$$\mathbf{y} = \Phi\mathbf{z} + \mathbf{w}^\perp, \quad (28)$$

where

$$\mathbf{z} = \mathbf{x}_{\text{true}} + \mathbf{v}. \quad (29)$$

We begin by computing the limit of the norms of the measurement vector \mathbf{y} and the projected noise vector \mathbf{w}^\perp .

Lemma 1: The limits

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{\|\mathbf{y}\|^2}{1 + \|\mathbf{x}\|^2} &= 1, \\ \lim_{n \rightarrow \infty} \|\mathbf{w}^\perp\|^2 &= 1, \end{aligned}$$

hold almost surely and in probability.

Proof: The vector \mathbf{w} is Gaussian, zero mean and white with variance $1/m$ per entry. Therefore, its projection, \mathbf{w}^\perp , will also be white in the $(m-k)$ -dimensional orthogonal complement of the range of Φ with variance $1/m$ per dimension. Therefore, by the strong law of large numbers

$$\lim_{n \rightarrow \infty} \|\mathbf{w}^\perp\|^2 = \lim_{n \rightarrow \infty} \frac{m-k}{m} = 1,$$

where the last step follows from the fact that (9) implies that $k/m \rightarrow 0$.

Similarly, it is easily verified that since \mathbf{A} and \mathbf{w} have i.i.d. Gaussian entries with variance $1/m$, the vector \mathbf{y} is also i.i.d. Gaussian with per-entry variance $(\|\mathbf{x}\|^2 + 1)/m$. Again, the strong law of large numbers shows that

$$\lim_{n \rightarrow \infty} \frac{\|\mathbf{y}\|^2}{1 + \|\mathbf{x}\|^2} = 1. \quad \blacksquare$$

We next need to compute the minimum singular value of Φ .

Lemma 2: Let $\sigma_{\min}(\Phi)$ and $\sigma_{\max}(\Phi)$ be the minimum and maximum singular values of Φ , respectively. Then

$$\lim_{n \rightarrow \infty} \sigma_{\min}(\Phi) = \lim_{n \rightarrow \infty} \sigma_{\max}(\Phi) = 1$$

where the limits are in probability.

Proof: Since the matrix Φ has $\mathcal{N}(0, 1/m)$ i.i.d. entries, the Marčenko–Pastur theorem [35] states that

$$\begin{aligned} \lim_{n \rightarrow \infty} \sigma_{\min}(\Phi) &= \lim_{n \rightarrow \infty} 1 - \sqrt{k/m} \\ \lim_{n \rightarrow \infty} \sigma_{\max}(\Phi) &= \lim_{n \rightarrow \infty} 1 + \sqrt{k/m} \end{aligned}$$

where the limits are in probability. The result now follows from (9) which implies that $k/m \rightarrow 0$ as $n \rightarrow \infty$. \blacksquare

We can also bound the singular values of submatrices of Φ . Given a subset $I \subseteq \{1, 2, \dots, k\}$, let Φ_I be the submatrix of Φ formed by the columns \mathbf{a}_i for $i \in I$. Also, let \mathbf{P}_I be the projection onto the orthogonal complement of the span of the set $\{\mathbf{a}_i, i \in I\}$. We have the following bound.

Lemma 3: Let I and J be any two disjoint subsets of indices such that

$$I \cup J = \{1, 2, \dots, k\}.$$

Then,

$$\sigma_{\min}(\Phi_I' \mathbf{P}_I \Phi_J) \geq \sigma_{\min}^2(\Phi).$$

Proof: The matrix $\mathbf{S} = [\Phi_I \ \Phi_J]$ is identical to Φ except that the columns may be permuted. In particular, $\sigma_{\min}(\mathbf{S}) =$

$\sigma_{\min}(\Phi)$. Therefore,

$$\begin{aligned} \mathbf{S}'\mathbf{S} &= \begin{bmatrix} \Phi'_I\Phi_I & \Phi'_I\Phi_J \\ \Phi'_J\Phi_I & \Phi'_J\Phi_J \end{bmatrix} \\ &\geq \sigma_{\min}^2(\mathbf{S})I \\ &= \sigma_{\min}^2(\Phi)I \\ &\geq \begin{bmatrix} 0 & 0 \\ 0 & \sigma_{\min}^2(\Phi)I \end{bmatrix}. \end{aligned}$$

The Schur complement (see, for example [36]) now shows that

$$\Phi'_J\Phi_J - \sigma_{\min}^2(\Phi)I \geq \Phi'_J\Phi_I(\Phi'_I\Phi_I)^{-1}\Phi'_I\Phi_J,$$

or equivalently,

$$\Phi'_J(I - \Phi_I(\Phi'_I\Phi_I)^{-1}\Phi'_I)\Phi_J \geq \sigma_{\min}^2(\Phi)I.$$

The result now follows from the fact that

$$\mathbf{P}_I = I - \Phi_I(\Phi'_I\Phi_I)^{-1}\Phi'_I.$$

■

We also need the following tail bound on chi-squared random variables.

Lemma 4: Suppose $X_i, i = 1, 2, \dots$, is a sequence of real-valued, scalar Gaussian random variables with $X_i \sim \mathcal{N}(0, 1)$. The variables need not be independent. Let M_k be the maximum

$$M_k = \max_{i=1, \dots, k} |X_i|^2.$$

Then

$$\limsup_{k \rightarrow \infty} \frac{M_k}{2 \log(k)} \leq 1,$$

where the limit is in probability.

Proof: See for example [28].

■

This bound permits us to bound the minimum component of \mathbf{z} .

Lemma 5: Let z_{\min} be the minimum component value

$$z_{\min} = \min_{j=1, \dots, k} |z_j|. \quad (30)$$

Then

$$\liminf_{n \rightarrow \infty} \frac{z_{\min}}{x_{\min}} \geq 1,$$

where the limit is in probability and x_{\min} is defined in (11).

Proof: Since \mathbf{w} is zero mean and Gaussian, so is \mathbf{v} as defined in (27). Also, the covariance of \mathbf{v} is bounded above by

$$\begin{aligned} \mathbf{E}[\mathbf{v}\mathbf{v}'] &\stackrel{(a)}{=} (\Phi'\Phi)^{-1}\Phi'(\mathbf{E}[\mathbf{w}\mathbf{w}'])\Phi'(\Phi'\Phi)^{-1} \\ &\stackrel{(b)}{=} \frac{1}{m}(\Phi'\Phi)^{-1} \\ &\stackrel{(c)}{\leq} \frac{1}{m}\sigma_{\min}^{-2}(\Phi), \end{aligned}$$

where (a) follows from the definition of \mathbf{v} in (27); (b) follows from the assumption that $\mathbf{E}[\mathbf{w}\mathbf{w}'] = (1/m)I_m$; and (c) is a basic property of singular values. This implies that for every $i \in \{1, 2, \dots, k\}$,

$$\mathbf{E}|v_i|^2 \leq \frac{1}{m}\sigma_{\min}^{-2}(\Phi).$$

Applying Lemma 4 shows that

$$\limsup_{k \rightarrow \infty} \frac{mv_{\max}^2\sigma_{\min}^2(\Phi)}{2 \log(k)} \leq 1, \quad (31)$$

where

$$v_{\max} = \max_{i=1, \dots, k} |v_i|.$$

Therefore,

$$\begin{aligned} &\lim_{n \rightarrow \infty} \frac{v_{\max}^2}{x_{\min}^2} \\ &= \lim_{n \rightarrow \infty} \left(\frac{mv_{\max}^2}{2 \log(k)} \right) \left(\frac{2 \log(k)}{mx_{\min}^2} \right) \\ &\stackrel{(a)}{\leq} \lim_{n \rightarrow \infty} \left(\frac{mv_{\max}^2\sigma_{\min}^2(\Phi)}{2 \log(k)} \right) \left(\frac{2 \log(k)}{mx_{\min}^2} \right) \\ &\stackrel{(b)}{\leq} \lim_{n \rightarrow \infty} \frac{2 \log(k)}{mx_{\min}^2} \\ &\stackrel{(c)}{\leq} \lim_{n \rightarrow \infty} \frac{2 \log(n-k)}{mx_{\min}^2} \\ &\stackrel{(d)}{\leq} \lim_{n \rightarrow \infty} \frac{1}{(1+\delta)kx_{\min}^2} \\ &\stackrel{(e)}{=} 0, \end{aligned}$$

where all the limits are in probability and (a) follows from Lemma 2; (b) follows from (31); (c) follows from the fact that $k < n/2$ and hence $k < n - k$; (d) follows from (9); and (e) follows from (10). Now, for $j \in \{1, 2, \dots, k\}$,

$$|z_j| = |x_j + v_j| \geq |x_j| - |v_j|,$$

and therefore,

$$z_{\min} \geq x_{\min} - v_{\max}.$$

Hence,

$$\frac{z_{\min}}{x_{\min}} \geq 1 - \frac{v_{\max}}{x_{\min}} \rightarrow 1,$$

where again the limit is in probability.

■

D. Probability of Missed Detection

With the bounds in the previous section, we can now show that the probability of missed detection goes to zero. The proof is similar to Tropp and Gilbert's proof in [10] with some modifications to account for the noise.

For any $t \in \{0, 1, \dots, k\}$, let $J(t) = I_{\text{true}} \cap I_{\text{true}}(t)^c$, which is the set of indices $j \in I_{\text{true}}$ that are *not* yet detected in iteration t of the genie algorithm in Section VIII-A. Then

$$\Phi\mathbf{z} = \Phi_{I_{\text{true}}(t)}\mathbf{z}_{I_{\text{true}}(t)} + \Phi_{J(t)}\mathbf{z}_{J(t)}, \quad (32)$$

where (using the notation of the previous subsection), Φ_I denotes the submatrix of Φ formed by the columns with indices $i \in I$, and \mathbf{z}_I denotes the corresponding subvector.

Now since $\mathbf{P}_{\text{true}}(t)$ is the projection onto the orthogonal complement of the span of $\{\mathbf{a}_i, i \in I_{\text{true}}(t)\}$,

$$\mathbf{P}_{\text{true}}(t)\Phi_{I_{\text{true}}(t)} = 0. \quad (33)$$

Also, since \mathbf{w}^\perp is orthogonal to \mathbf{a}_i for all $i \in I_{\text{true}}$ and $I_{\text{true}}(t) \subseteq I_{\text{true}}$,

$$\mathbf{P}_{\text{true}}(t)\mathbf{w}^\perp = \mathbf{w}^\perp. \quad (34)$$

Therefore,

$$\begin{aligned} \mathbf{P}_{\text{true}}(t)\mathbf{y} &\stackrel{(a)}{=} \mathbf{P}_{\text{true}}(t)(\Phi\mathbf{z} + \mathbf{w}^\perp) \\ &\stackrel{(b)}{=} \mathbf{P}_{\text{true}}(t)(\Phi_{J(t)}\mathbf{z}_{J(t)} + \mathbf{w}^\perp) \\ &\stackrel{(c)}{=} \mathbf{P}_{\text{true}}(t)\Phi_{J(t)}\mathbf{z}_{J(t)} + \mathbf{w}^\perp, \end{aligned} \quad (35)$$

where (a) follows from (28); (b) follows from (32) and (33); and (c) follows from (34).

Now using (34) and the fact that \mathbf{w}^\perp is orthogonal to \mathbf{a}_i for all $i \in I_{\text{true}}$, we have

$$\mathbf{a}_i' \mathbf{P}_{\text{true}}(t)\mathbf{w}^\perp = \mathbf{a}_i' \mathbf{w}^\perp = 0 \quad (36)$$

for all $i \in I_{\text{true}}$. Since the columns of $\Phi_{J(t)}$ are formed by vectors \mathbf{a}_i with $i \in I_{\text{true}}$,

$$\Phi_{J(t)}' \mathbf{P}_{\text{true}}(t)\mathbf{w}^\perp = 0. \quad (37)$$

Combining (37) and (35),

$$\|\mathbf{P}_{\text{true}}(t)\mathbf{y}\|^2 = \|\mathbf{P}_{\text{true}}(t)\Phi_{J(t)}\mathbf{z}_{J(t)}\|^2 + \|\mathbf{w}^\perp\|^2. \quad (38)$$

Now for all t , we have that

$$\begin{aligned} &\max_{j \in I_{\text{true}}} \rho_{\text{true}}(t, j) \\ &\stackrel{(a)}{=} \frac{1}{\|\mathbf{P}_{\text{true}}(t)\mathbf{y}\|^2} \max_{j \in I_{\text{true}}} |\mathbf{a}_j' \mathbf{P}_{\text{true}}(j)\mathbf{y}|^2 \\ &\stackrel{(b)}{=} \frac{1}{\|\mathbf{P}_{\text{true}}(t)\mathbf{y}\|^2} \max_{j \in J(t)} |\mathbf{a}_j' \mathbf{P}_{\text{true}}(j)\mathbf{y}|^2 \\ &\stackrel{(c)}{=} \frac{1}{\|\mathbf{P}_{\text{true}}(t)\mathbf{y}\|^2} \|\Phi_{J(t)}' \mathbf{P}_{\text{true}}(j)\mathbf{y}\|_\infty^2 \\ &\stackrel{(d)}{\geq} \frac{1}{|J(t)| \|\mathbf{P}_{\text{true}}(t)\mathbf{y}\|^2} \|\Phi_{J(t)}' \mathbf{P}_{\text{true}}(j)\mathbf{y}\|_2^2 \\ &\stackrel{(e)}{=} \frac{\|\Phi_{J(t)}' \mathbf{P}_{\text{true}}(j)\Phi_{J(t)}\mathbf{z}_{J(t)}\|_2^2}{|J(t)| \|\mathbf{P}_{\text{true}}(t)\mathbf{y}\|^2} \\ &\stackrel{(f)}{=} \frac{\|\Phi_{J(t)}' \mathbf{P}_{\text{true}}(j)\Phi_{J(t)}\mathbf{z}_{J(t)}\|_2^2}{|J(t)| (\|\mathbf{P}_{\text{true}}(t)\Phi_{J(t)}\mathbf{z}_{J(t)}\|^2 + \|\mathbf{w}^\perp\|^2)} \\ &\stackrel{(g)}{\geq} \frac{\sigma_{\min}(\Phi_{J(t)}' \mathbf{P}_{\text{true}}(j)\Phi_{J(t)}) \|\mathbf{z}_{J(t)}\|_2^2}{|J(t)| (\sigma_{\max}^2(\Phi) \|\mathbf{z}_{J(t)}\|^2 + \|\mathbf{w}^\perp\|^2)} \\ &\stackrel{(h)}{\geq} \frac{\sigma_{\min}^4(\Phi) \|\mathbf{z}_{J(t)}\|_2^2}{|J(t)| (\sigma_{\max}^2(\Phi) \|\mathbf{z}_{J(t)}\|^2 + \|\mathbf{w}^\perp\|^2)} \\ &\stackrel{(i)}{\geq} \frac{\sigma_{\min}^4(\Phi) z_{\min}^2}{\sigma_{\max}^2(\Phi) k z_{\min}^2 + \|\mathbf{w}^\perp\|^2}, \end{aligned} \quad (39)$$

where (a) follows from the definition of $\rho_{\text{true}}(t, j)$ in (16); (b) follows from the fact that $\mathbf{P}_{\text{true}}(t)\mathbf{a}_j = 0$ for all $j \in I_{\text{true}}(t)$ and hence the maximum will occur on the set $j \in I_{\text{true}} \cap I_{\text{true}}(t)^c = J(t)$; (c) follows from the fact that $\Phi_{J(t)}$ is the matrix of the columns \mathbf{a}_j with $j \in J(t)$; (d) follows the bound that $\|\mathbf{v}\|_2^2 \leq d \|\mathbf{v}\|_\infty^2$ for any $\mathbf{v} \in \mathbb{R}^d$; (e) follows (35) and (37); (f) follows from (38); (g) follows from the fact that $\mathbf{P}_{\text{true}}(t)$ is a projection operator and hence,

$$\sigma_{\max}(\mathbf{P}_{\text{true}}(t)\Phi_{J(t)}) \leq \sigma_{\max}(\Phi_{J(t)}) \leq \sigma_{\max}(\Phi);$$

(h) follows from Lemma 3; and (i) follows from the bound

$$\|\mathbf{z}_{J(t)}\|^2 \geq |J(t)| z_{\min}^2$$

and $|J(t)| \leq k$. Therefore,

$$\begin{aligned} &\liminf_{n \rightarrow \infty} \min_{t=0, \dots, k-1} \max_{j \in I_{\text{true}}} \frac{1}{\mu} \rho_{\text{true}}(t, j) \\ &\stackrel{(a)}{\geq} \liminf_{n \rightarrow \infty} \frac{1}{\mu} \frac{\sigma_{\min}^4(\Phi) z_{\min}^2}{\sigma_{\max}^2(\Phi) k z_{\min}^2 + \|\mathbf{w}^\perp\|^2} \\ &\stackrel{(b)}{\geq} \liminf_{n \rightarrow \infty} \frac{1}{\mu} \frac{z_{\min}^2}{k z_{\min}^2 + 1} \\ &\stackrel{(c)}{\geq} \liminf_{n \rightarrow \infty} \frac{1}{\mu} \frac{x_{\min}^2}{k x_{\min}^2 + 1} \\ &\stackrel{(d)}{\geq} \liminf_{n \rightarrow \infty} \frac{1}{k\mu} \\ &\stackrel{(e)}{\geq} 1 + \epsilon, \end{aligned} \quad (40)$$

where (a) follows from (39), (b) follows from Lemmas 1 and 2; (c) follows from Lemma 5; (d) follows from the assumption of the theorem that $kx_{\min}^2 \rightarrow \infty$; and (e) follows from (24). The definition of p_{MD} in (19) now shows that

$$\lim_{n \rightarrow \infty} p_{\text{MD}} = 0.$$

E. Bounds on Normalized Brownian Motions

Let $B(t)$ be a standard Brownian motion. Define the *normalized Brownian motion* $S(t)$ as the process

$$S(t) = \frac{1}{\sqrt{t}} B(t), \quad t > 0. \quad (41)$$

We call the process normalized since

$$\mathbf{E}|S(t)|^2 = \frac{1}{t} \mathbf{E}|B(t)|^2 = \frac{t}{t} = 1.$$

We first characterize the autocorrelation of this process.

Lemma 6: If $t > s$, the normalized Brownian motion has autocorrelation

$$\mathbf{E}[S(t)S(s)] = \sqrt{s/t}.$$

Proof: Write

$$S(t) = \frac{1}{\sqrt{t}} (B(s) + (B(t) - B(s))).$$

Thus,

$$\begin{aligned} \mathbf{E}[S(t)S(s)] &= \frac{1}{\sqrt{st}} \mathbf{E}[(B(s) + (B(t) - B(s)))B(s)] \\ &\stackrel{(a)}{=} \frac{1}{\sqrt{st}} \mathbf{E}[B(s)^2] \\ &\stackrel{(b)}{=} \frac{s}{\sqrt{st}} = \sqrt{\frac{s}{t}}, \end{aligned}$$

where (a) follows from the orthogonal increments property of Brownian motions; and (b) follows from the fact that $B(s) \sim \mathcal{N}(0, s)$. ■

We now need the following standard Gaussian tail bound.

Lemma 7: Suppose X is a real-valued, scalar Gaussian random variable, $X \sim \mathcal{N}(0, 1)$. Then,

$$\Pr(X^2 > \mu) \leq \frac{1}{\sqrt{\pi\mu}} \exp(-\mu/2).$$

Proof: See for example [37]. ■

We next provide a simple bound on the maximum of sample paths of $S(t)$.

Lemma 8: For any $0 < a < b$, let

$$S_{\max}(a, b) = \sup_{t \in [a, b]} |S(t)|.$$

Then, for any $\mu > 0$,

$$\Pr(S_{\max}^2(a, b) > \mu) \leq \frac{2b}{a\mu\sqrt{\pi}} \exp\left(-\frac{a\mu}{2b}\right).$$

Proof: Since $S(t)$ and $S(-t)$ are identically distributed,

$$\Pr(S_{\max}^2(a, b) > \mu) \leq 2 \Pr\left(\sup_{t \in [a, b]} S(t) > \sqrt{\mu}\right). \quad (42)$$

So, it will suffice to bound the probability of the single-sided event $\sup S(t) > \sqrt{\mu}$. For $t \geq 0$, define $B_a(t) = B(a+t) - B(a)$. Then, $B_a(t)$ is a standard Brownian motion independent of $B(a)$. Also,

$$\begin{aligned} & \sup_{t \in [a, b]} S(t) > \sqrt{\mu} \\ \Rightarrow & \sup_{t \in [a, b]} \frac{1}{\sqrt{t}} B(t) > \sqrt{\mu} \\ \Rightarrow & \sup_{t \in [a, b]} B(t) > \sqrt{a\mu} \\ \Rightarrow & B(a) + \sup_{t \in [0, b-a]} B_a(t) > \sqrt{a\mu}. \end{aligned}$$

Now, the reflection principle (see, for example [38]) states that for any y ,

$$\Pr\left(\max_{t \in [0, b-a]} B_a(t) > y\right) = 2 \Pr\left(\sqrt{b-a}Y > y\right),$$

where Y is a unit-variance, zero-mean Gaussian. Also, $B(a) \sim \mathcal{N}(0, a)$, so if we define $X = (1/\sqrt{a})B(a)$, then $X \sim \mathcal{N}(0, 1)$. Since $B(a)$ is independent of $B_a(t)$ for all $t \geq 0$, we can write

$$\begin{aligned} & \Pr\left(\sup_{t \in [a, b]} S(t) > \sqrt{\mu}\right) \\ & \leq 2 \Pr\left(\sqrt{a}X + \sqrt{b-a}Y > \sqrt{a\mu}\right), \end{aligned} \quad (43)$$

where X and Y are independent zero mean Gaussian random variables with unit variance. Now $\sqrt{a}X + \sqrt{b-a}Y$ has variance

$$\mathbb{E}\left[(\sqrt{a}X + \sqrt{b-a}Y)^2\right] = a + b - a = b.$$

Applying Lemma 7 shows that (43) can be bounded by

$$\Pr\left(\sup_{t \in [a, b]} S(t) > \sqrt{\mu}\right) \leq \frac{b}{a\mu\sqrt{\pi}} \exp\left(-\frac{a\mu}{2b}\right).$$

Substituting this bound in (42) proves the lemma. \blacksquare

Our next lemma improves the bound for large μ .

Lemma 9: There exist constants C_1 , C_2 , and C_3 such that for any $0 < a < b$ and $\mu > C_3$,

$$\Pr(S_{\max}^2(a, b) > \mu) \leq (C_1 + C_2 \log(b/a)) e^{-\mu/2}.$$

Proof: Fix any integer $n > 0$, and define $t_i = a(b/a)^{i/n}$ for $i = 0, 1, \dots, n$. Observe that t_i s partition the interval $[a, b]$ in that

$$a = t_0 < t_1 < \dots < t_n = b.$$

Also, let $r = b/a$. Then, $t_{i+1}/t_i = (b/a)^{1/n} = r^{1/n}$. Applying Lemma 8 to each interval in the partition,

$$\begin{aligned} & \Pr(S_{\max}^2(a, b) > \mu) \\ & \leq \sum_{i=1}^{n-1} \Pr(S_{\max}^2(t_i, t_{i+1}) > \mu) \\ & \leq \frac{nr^{1/n}}{\mu\sqrt{\pi}} \exp\left(-\frac{r^{-1/n}\mu}{2}\right). \end{aligned} \quad (44)$$

Now, let $\delta > 0$, and for $\mu > \delta$, let

$$n = \left\lceil -\frac{\log(r)}{\log(1 - \delta/\mu)} \right\rceil. \quad (45)$$

Then

$$r^{-1/n} \geq 1 - \delta/\mu, \quad (46)$$

and hence

$$\exp\left(-\frac{r^{-1/n}\mu}{2}\right) \leq e^{\delta/2} e^{-\mu/2}. \quad (47)$$

Also, (45) implies that

$$n \leq 1 - \frac{\log(r)}{\log(1 - \delta/\mu)} \leq 1 + \frac{\mu}{\delta} \log(r), \quad (48)$$

where we have used the fact that $\log(1 - x) < -x$ for $x > 0$. Combining the bounds (46) and (48) yields

$$\frac{nr^{1/n}}{\mu} \leq \left(1 + \frac{\mu}{\delta} \log(r)\right) \frac{1}{\mu - \delta}. \quad (49)$$

Now, pick any $\delta > 0$ and let $C_3 = 2\delta$. Then if $\mu > C_3 = 2\delta$, (49) implies that

$$\frac{nr^{1/n}}{\mu^2} \leq \frac{1}{\delta} (1 + 2 \log(r)). \quad (50)$$

Substituting (47) and (50) into (44) shows that

$$\Pr(S_{\max}(a, b) > \mu) \leq (C_1 + C_2 \log(r)) e^{-\mu/2},$$

where

$$C_1 = \frac{e^{\delta/2}}{\sqrt{\pi}\delta}, \quad C_2 = \frac{2e^{\delta/2}}{\sqrt{\pi}\delta}.$$

The result now follows from the fact that $r = b/a$. \blacksquare

F. Bounds on Sequences of Projections

We can now apply the results in the previous subsection to bound the norms of sequences of projections. Let $\mathbf{y} \in \mathbb{R}^m$ be any deterministic vector, and let $\mathbf{P}(i)$, $i = 0, 1, \dots, k$ be a deterministic sequence of orthogonal projection operators on \mathbb{R}^m . Assume that the sequence $\mathbf{P}(i)$ is *decreasing* in that $\mathbf{P}(i)\mathbf{P}(j) = \mathbf{P}(i)$ for $j > i$.

Lemma 10: Let $\mathbf{a} \in \mathbb{R}^m$ be a Gaussian random vector with unit variance, and define the random variable

$$M = \max_{i=0, \dots, k} \frac{|\mathbf{a}^T \mathbf{P}(i) \mathbf{y}|^2}{\|\mathbf{P}(i) \mathbf{y}\|^2}.$$

Then there exist constants C_1 , C_2 , and $C_3 > 0$ (all independent of the problem parameters) such that $\mu > C_3$ implies

$$\Pr(M > \mu) \leq (C_1 + C_2 \log(r)) e^{-\mu/2},$$

where $r = \|\mathbf{P}(1)\mathbf{y}\|^2 / \|\mathbf{P}(n)\mathbf{y}\|^2$.

Proof: Define

$$z_i = \frac{\mathbf{y}'\mathbf{P}(i)\mathbf{a}}{\|\mathbf{P}(i)\mathbf{y}\|},$$

so that

$$M = \max_{i=0,\dots,k} |z_i|^2.$$

Since each z_i is the inner product of the Gaussian vector \mathbf{a} with a fixed vector, the scalars $\{z_i, i = 0, 1, \dots, k\}$ are jointly Gaussian. Since \mathbf{a} has mean zero, so do the z_i s.

To compute the cross-correlations, suppose that $j \geq i$. Then

$$\begin{aligned} \mathbf{E}[z_i \bar{z}_j] &= \frac{1}{\|\mathbf{P}(i)\mathbf{y}\| \|\mathbf{P}(j)\mathbf{y}\|} \mathbf{E}[\mathbf{y}'\mathbf{P}(i)\mathbf{a}\mathbf{a}'\mathbf{P}(j)\mathbf{y}] \\ &\stackrel{(a)}{=} \frac{1}{\|\mathbf{P}(i)\mathbf{y}\| \|\mathbf{P}(j)\mathbf{y}\|} \mathbf{y}'\mathbf{P}(i)\mathbf{P}(j)\mathbf{y} \\ &\stackrel{(b)}{=} \frac{1}{\|\mathbf{P}(i)\mathbf{y}\| \|\mathbf{P}(j)\mathbf{y}\|} \mathbf{y}'\mathbf{P}(i)\mathbf{y} \\ &= \frac{\|\mathbf{P}(i)\mathbf{y}\|}{\|\mathbf{P}(j)\mathbf{y}\|}, \end{aligned}$$

where (a) uses the fact that $\mathbf{E}[\mathbf{a}\mathbf{a}'] = I_m$; and (b) uses the descending property that $\mathbf{P}(i)\mathbf{P}(j) = \mathbf{P}(i)$. Therefore, if we let $t_i = \|\mathbf{P}(i)\mathbf{y}\|^2$, we have the cross-correlations

$$\mathbf{E}[z_i \bar{z}_j] = \sqrt{t_i/t_j} \quad (51)$$

for all $j \geq i$. Also observe that since the projection operators are decreasing, so are the t_j s. That is, for $j \geq i$,

$$t_i = \|\mathbf{P}(i)\mathbf{y}\|^2 \stackrel{(a)}{=} \|\mathbf{P}(i)\mathbf{P}(j)\mathbf{y}\|^2 \stackrel{(b)}{\leq} \|\mathbf{P}(j)\mathbf{y}\|^2 = t_j,$$

where again (a) uses the decreasing property; and (b) uses the fact that $\mathbf{P}(i)$ is a projection operator and norm non-increasing.

Now let $S(t)$ be the normalized Brownian motion in (41). Lemma 6 and (51) show that the Gaussian vector

$$\mathbf{z} = (z_0, z_1, \dots, z_k)$$

has the same covariance as the vector of samples of $S(t)$,

$$\mathbf{s} = (S(t_0), S(t_1), \dots, S(t_k)).$$

Since they are also both zero-mean and Gaussian, they have the same distribution. Hence, for all μ ,

$$\begin{aligned} \Pr(M > \mu) &= \Pr\left(\max_{i=0,\dots,k} |z_i|^2 > \mu\right) \\ &= \Pr\left(\max_{i=0,\dots,k} |S(t_i)|^2 > \mu\right) \\ &\leq \Pr\left(\sup_{t \in [t_k, t_0]} |S(t)|^2 > \mu\right), \end{aligned}$$

where the last step follows from the fact that the t_i s are decreasing and hence $t_k \geq t_i \geq t_0$ for all $i \in \{0, 1, \dots, k\}$. The result now follows from Lemma 9. ■

G. Probability of False Alarm

Recall that all the projection operators $\mathbf{P}_{\text{true}}(t)$ and the vector \mathbf{y} are statistically independent of the vectors \mathbf{a}_j for $j \notin I_{\text{true}}$. Since the entries of the matrix \mathbf{A} are i.i.d. Gaussian with zero mean and variance $1/m$, the vector $m\mathbf{a}_j$ is Gaussian with unit variance. Hence, Lemma 10 shows that there exist constants C_1 , C_2 , and C_3 such that for any $\lambda > C_3$,

$$\Pr\left(\max_{t=0,\dots,k} m \frac{|\mathbf{a}_j \mathbf{P}_{\text{true}}(t)\mathbf{y}|^2}{\|\mathbf{P}_{\text{true}}(t)\mathbf{y}\|^2} \geq \lambda\right) \leq B e^{-\lambda/2}, \quad (52)$$

where $j \notin I_{\text{true}}$ and

$$B = C_1 + C_2 \log\left(\frac{\|\mathbf{P}_{\text{true}}(0)\mathbf{y}\|^2}{\|\mathbf{P}_{\text{true}}(k)\mathbf{y}\|^2}\right). \quad (53)$$

Therefore,

$$\begin{aligned} p_{\text{FA}} &\stackrel{(a)}{=} \Pr\left(\max_{t=1,\dots,k} \max_{j \notin I_{\text{true}}} \rho_{\text{true}}(t, j) > \mu\right) \\ &\stackrel{(b)}{\leq} (n-k) \max_{j \notin I_{\text{true}}} \Pr\left(\max_{t=1,\dots,k} \rho_{\text{true}}(t, j) > \mu\right) \\ &\stackrel{(c)}{=} (n-k) \max_{j \notin I_{\text{true}}} \Pr\left(\max_{t=1,\dots,k} \frac{|\mathbf{a}_j \mathbf{P}_{\text{true}}(t)\mathbf{y}|^2}{\|\mathbf{P}_{\text{true}}(t)\mathbf{y}\|^2} > \mu\right) \\ &\stackrel{(d)}{\leq} (n-k) B e^{-\mu/2} \\ &\stackrel{(e)}{\leq} (n-k) B e^{-(1+\epsilon)\log(n-k)} \\ &= \frac{1}{(n-k)^\epsilon} B, \end{aligned} \quad (54)$$

where (a) follows from the definition of p_{FA} in (20); (b) uses the union bound and the fact that I_{true}^c has $n-k$ elements; (c) follows from the definition of $\rho_{\text{true}}(t, j)$ in (16); (d) follows from (52) under the condition that $\mu m > C_3$; and (e) follows from (23). By (9) and the hypothesis of the theorem that $n-k \rightarrow \infty$,

$$\mu m = (1+\delta)2\log(n-k) \rightarrow \infty \text{ as } n \rightarrow \infty.$$

Therefore, for sufficiently large n , $\mu m > C_3$ and (54) holds.

Now, since $I_{\text{true}}(0) = \emptyset$, $\mathbf{P}_{\text{true}}(0) = I$ and therefore

$$\mathbf{P}_{\text{true}}(0)\mathbf{y} = \mathbf{y}. \quad (55)$$

Also, $I_{\text{true}}(k) = I_{\text{true}}$ and so $\mathbf{P}_{\text{true}}(k)$ is the projection onto the orthogonal complement of the range of Φ . Hence $\mathbf{P}_{\text{true}}(k)\Phi = 0$. Combining this fact with (28) and (34) shows

$$\mathbf{P}_{\text{true}}(k)\mathbf{y} = \mathbf{w}^\perp. \quad (56)$$

Therefore,

$$\begin{aligned} \liminf_{n \rightarrow \infty} p_{\text{FA}} &\stackrel{(a)}{\leq} \liminf_{n \rightarrow \infty} \frac{1}{(n-k)^\epsilon} B \\ &\stackrel{(b)}{\leq} \liminf_{n \rightarrow \infty} \frac{1}{(n-k)^\epsilon} \left(C_1 + C_2 \log\left(\frac{\|\mathbf{P}_{\text{true}}(0)\mathbf{y}\|^2}{\|\mathbf{P}_{\text{true}}(k)\mathbf{y}\|^2}\right)\right) \\ &\stackrel{(c)}{=} \liminf_{n \rightarrow \infty} \frac{1}{(n-k)^\epsilon} \left(C_1 + C_2 \log\left(\frac{\|\mathbf{y}\|^2}{\|\mathbf{w}^\perp\|^2}\right)\right) \\ &\stackrel{(d)}{=} \liminf_{n \rightarrow \infty} \frac{1}{(n-k)^\epsilon} (C_1 + C_2 \log(1 + \|\mathbf{x}\|^2)) \\ &\stackrel{(e)}{=} 0 \end{aligned}$$

where (a) follows from (54); (b) follows from (53); (c) follows from (55) and (56); (d) follows from Lemma 1; and (e) follows from (12). This completes the proof of the theorem.

ACKNOWLEDGMENTS

The authors thank Vivek Goyal for comments on an earlier draft and Martin Vetterli for his support, wisdom, and encouragement.

REFERENCES

- [1] S. Mallat, *A Wavelet Tour of Signal Processing*, 2nd ed. Academic Press, 1999.
- [2] A. Miller, *Subset Selection in Regression*, 2nd ed., ser. Monographs on Statistics and Applied Probability. New York: Chapman & Hall/CRC, 2002, no. 95.
- [3] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inform. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.
- [4] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inform. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.
- [5] E. J. Candès and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?" *IEEE Trans. Inform. Theory*, vol. 52, no. 12, pp. 5406–5425, Dec. 2006.
- [6] B. K. Natarajan, "Sparse approximate solutions to linear systems," *SIAM J. Computing*, vol. 24, no. 2, pp. 227–234, Apr. 1995.
- [7] S. Chen, S. A. Billings, and W. Luo, "Orthogonal least squares methods and their application to non-linear system identification," *Int. J. Control*, vol. 50, no. 5, pp. 1873–1896, Nov. 1989.
- [8] Y. C. Pati, R. Rezaifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Conf. Rec. 27th Asilomar Conf. Sig., Sys., & Comput.*, vol. 1, Pacific Grove, CA, Nov. 1993, pp. 40–44.
- [9] G. Davis, S. Mallat, and Z. Zhang, "Adaptive time-frequency decomposition," *Optical Eng.*, vol. 33, no. 7, pp. 2183–2191, Jul. 1994.
- [10] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Inform. Theory*, vol. 53, no. 12, pp. 4655–4666, Dec. 2007.
- [11] —, "Signal recovery from random measurements via orthogonal matching pursuit: The Gaussian case," California Inst. of Tech., Appl. Comput. Math. 2007-01, Aug. 2007.
- [12] J. A. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Trans. Inform. Theory*, vol. 50, no. 10, pp. 2231–2242, Oct. 2004.
- [13] D. L. Donoho, M. Elad, and V. N. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Trans. Inform. Theory*, vol. 52, no. 1, pp. 6–18, Jan. 2006.
- [14] M. A. Davenport and M. B. Wakin, "Analysis of orthogonal matching pursuit using the restricted isometry property," *IEEE Trans. Inform. Theory*, vol. 56, no. 9, pp. 4395–4401, Sep. 2010.
- [15] M. J. Wainwright, "Sharp thresholds for high-dimensional and noisy recovery of sparsity," Univ. of California, Berkeley, Dept. of Statistics, Tech. Rep., May 2006, arXiv:math.ST/0605740 v1 30 May 2006.
- [16] —, "Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1 -constrained quadratic programming (lasso)," *IEEE Trans. Inform. Theory*, vol. 55, no. 5, pp. 2183–2202, May 2009.
- [17] D. L. Donoho and J. Tanner, "Counting faces of randomly-projected polytopes when the projection radically lowers dimension," *J. Amer. Math. Soc.*, vol. 22, no. 1, pp. 1–53, Jan. 2009.
- [18] D. Needell and J. A. Tropp, "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples," *Appl. Comput. Harm. Anal.*, vol. 26, no. 3, pp. 301–321, May 2009.
- [19] W. Dai and O. Milenkovic, "Subspace pursuit for compressive sensing signal reconstruction," *IEEE Trans. Inform. Theory*, vol. 55, no. 5, pp. 2230–2249, May 2009.
- [20] D. L. Donoho, Y. Tsaig, I. Drori, and J. L. Starck, "Sparse solution of underdetermined linear equations by stagewise orthogonal matching pursuit," preprint, Mar. 2006.
- [21] D. Needell and R. Vershynin, "Uniform uncertainty principle and signal recovery via regularized orthogonal matching pursuit," *Found. Comput. Math.*, vol. 9, no. 3, pp. 317–334, Jun. 2009.
- [22] —, "Signal recovery from incomplete and inaccurate measurements via regularized orthogonal matching pursuit," *IEEE J. Sel. Topics Signal Process.*, vol. 4, no. 2, pp. 310–316, Apr. 2010.
- [23] A. K. Fletcher and S. Rangan, "Orthogonal matching pursuit from noisy measurements: A new analysis," in *Proc. Neural Information Process. Syst.*, Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, Eds., Vancouver, Canada, Dec. 2009.
- [24] M. J. Wainwright, "Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting," *IEEE Trans. Inform. Theory*, vol. 55, no. 12, pp. 5728–5741, Dec. 2009.
- [25] A. K. Fletcher, S. Rangan, and V. K. Goyal, "Necessary and sufficient conditions for sparsity pattern recovery," *IEEE Trans. Inform. Theory*, vol. 55, no. 12, pp. 5758–5772, Dec. 2009.
- [26] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Royal Stat. Soc., Ser. B*, vol. 58, no. 1, pp. 267–288, 1996.
- [27] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comp.*, vol. 20, no. 1, pp. 33–61, 1999.
- [28] A. K. Fletcher, S. Rangan, and V. K. Goyal, "On-off random access channels: A compressed sensing framework," arXiv:0903.1022v1 [cs.IT], Mar. 2009.
- [29] —, "A sparsity detection framework for on-off random access channels," in *Proc. IEEE Int. Symp. Inform. Theory*, Seoul, Korea, Jun.–Jul. 2009, pp. 169–173.
- [30] D. Wipf and B. Rao, "Comparing the effects of different weight distributions on finding sparse representations," in *Proc. Neural Information Process. Syst.*, Vancouver, Canada, Dec. 2006.
- [31] Y. Jin and B. Rao, "Performance limits of matching pursuit algorithms," in *Proc. IEEE Int. Symp. Inform. Theory*, Toronto, Canada, Jun. 2008, pp. 2444–2448.
- [32] M. Akçakaya and V. Tarokh, "Shannon-theoretic limits on noisy compressive sampling," *IEEE Trans. Inform. Theory*, vol. 56, no. 1, pp. 492–504, Jan. 2010.
- [33] G. Reeves, "Sparse signal sampling using noisy linear projections," Univ. of California, Berkeley, Dept. of Elec. Eng. and Comp. Sci., Tech. Rep. UCB/EECS-2008-3, Jan. 2008.
- [34] S. Aeron, V. Saligrama, and M. Zhao, "Information theoretic bounds for compressed sensing," *IEEE Trans. Inform. Theory*, vol. 56, no. 10, pp. 5111–5130, Oct. 2010.
- [35] V. A. Marčenko and L. A. Pastur, "Distribution of eigenvalues for some sets of random matrices," *Math. USSR-Sbornik*, vol. 1, no. 4, pp. 457–483, 1967.
- [36] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge Univ. Press, 1985, reprinted with corrections 1987.
- [37] J. Spanier and K. B. Oldham, *An Atlas of Functions*. Washington: Hemisphere Publishing, 1987.
- [38] I. Karatzas and S. E. Shreve, *Brownian Motion and Stochastic Calculus*, 2nd ed. New York, NY: Springer-Verlag, 1991.